Methodology
Lesson 5: Analysing our results

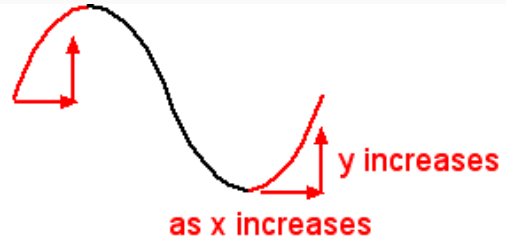
Carlos Ramisch
first.last@lis-lab.fr

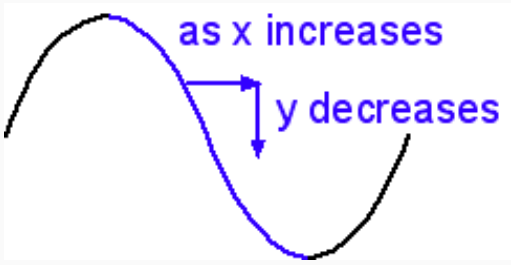
M2 IAAA - based on the course *Zen Research*
By Carlos Ramisch and Manon Scholivet

Outline

Two random variables

- For the moment we looked at random variables one by one
- It may be interesting to look at two random variables X and Y
  - $\rightarrow$ They may influence each other
  - $\rightarrow$ They may be both influenced by similar factors
- How does X and Y vary together?





Two variables: scatter plot

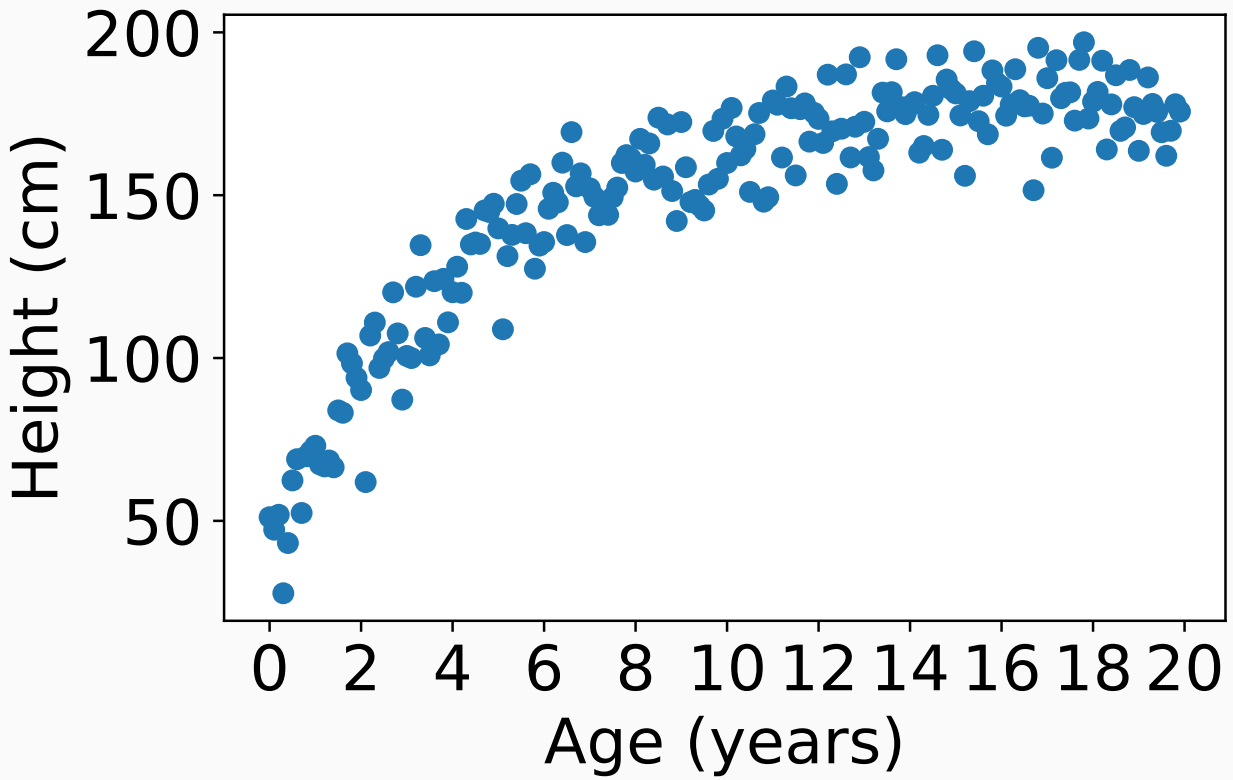
- Variable X on x-axis, variable Y on y-axis
- `plt.scatter(x,y)`
- The two variables are paired or aligned
  - $\rightarrow$ The sample consists of pairs of values
  - $\rightarrow$ Each value of X has a corresponding value of Y
  - $\rightarrow$ Both variables are numeric

Scatter plot example 1

A person's age (X) vs. height (Y)

Scatter plot example 1

A person's age (X) vs. height (Y)

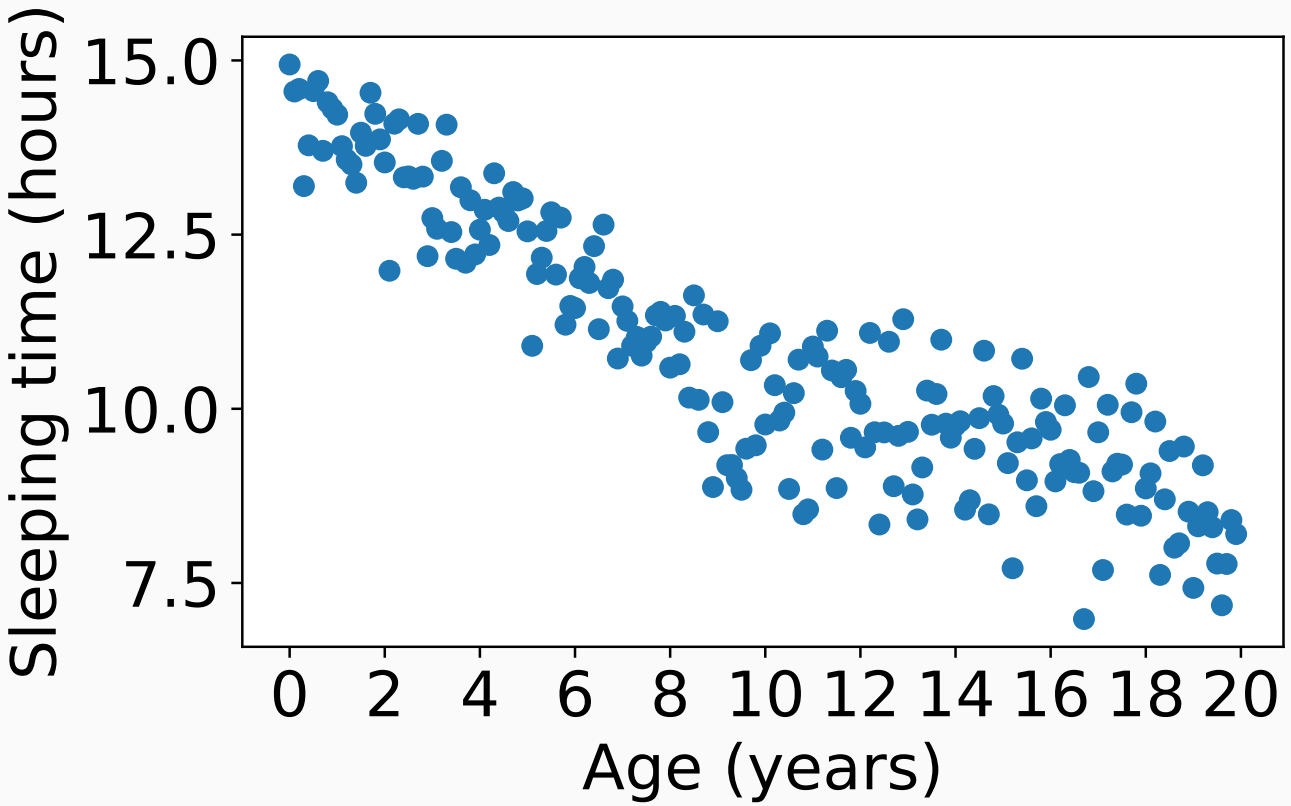


Scatter plot example 2

A person's age (X) vs. number of sleeping hours (Y)

Scatter plot example 2

A person's age (X) vs. number of sleeping hours (Y)

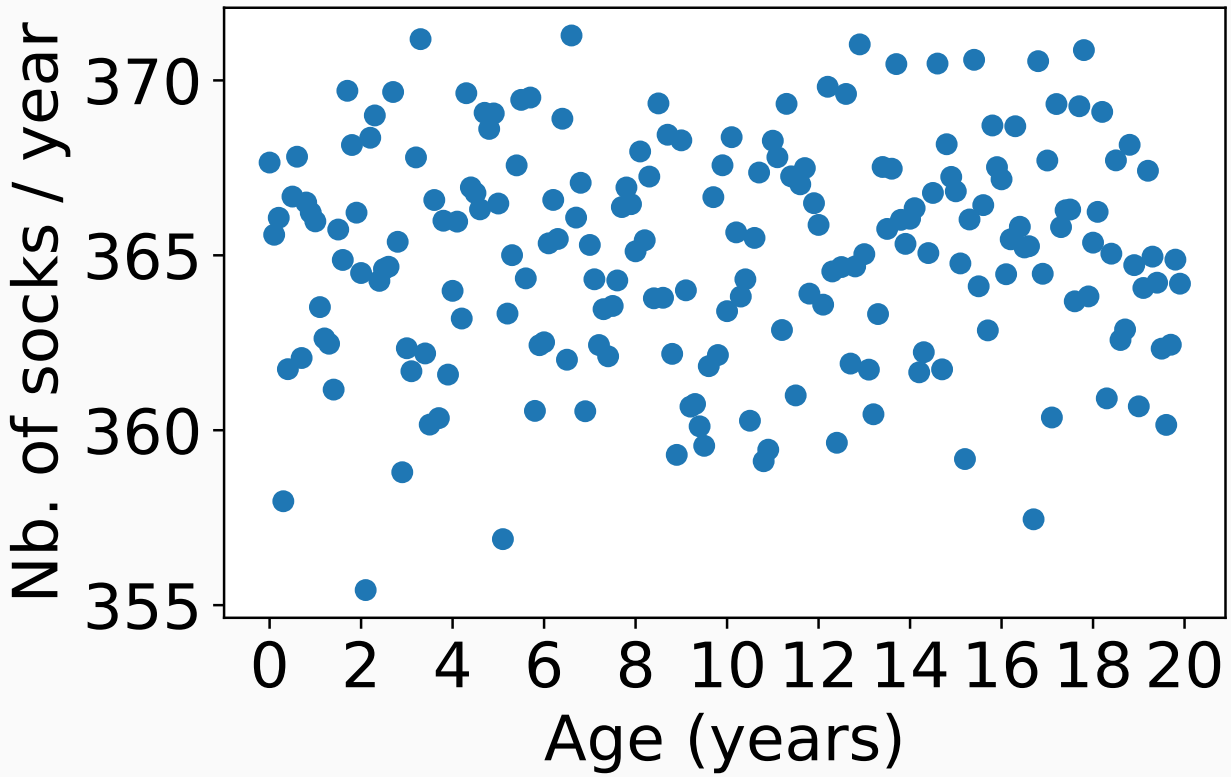


Scatter plot example 3

A person's age (X) vs. number of socks (Y) used per year (Y)

Scatter plot example 3

A person's age (X) vs. number of socks (Y) used per year (Y)



Example: compositionality and number of occurrences

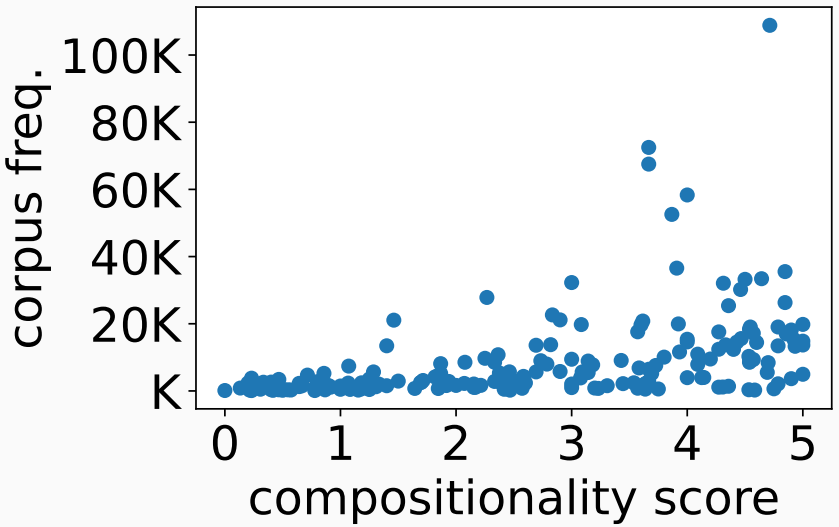
Jupyter notebook 6 & 7

- Hypothesis: frequent compounds are less compositional
- What is the relation between compositionality and frequency?

Example: compositionality and number of occurrences

Jupyter notebook 6 & 7

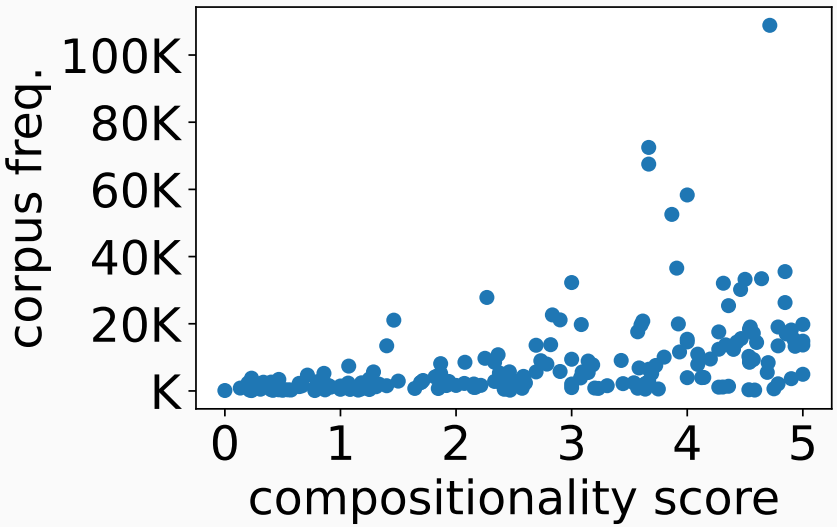
- Hypothesis: frequent compounds are less compositional
- What is the relation between compositionality and frequency?



Example: compositionality and number of occurrences

Jupyter notebook 6 & 7

- Hypothesis: frequent compounds are less compositional
- What is the relation between compositionality and frequency?



- Is there really something to see or are we over-interpreting?

Quantifying relations

- It would be nice to be able to quantify the relation!

Quantifying relations

- It would be nice to be able to quantify the relation!

We will obtain such metric in two steps:

1. Covariance
   - $\longrightarrow$ Not so easy to interpret
   - $\longrightarrow$ Computational step towards calculating correlation
2. Correlation
   - $\longrightarrow$ Much easier to interpret

Covariance: far from the mean

- Relation between each value x_i and the mean $\overline{x}$
- Relation between each value y_i and the mean $\overline{y}$

Covariance: far from the mean

- Relation between each value x_i and the mean $\overline{x}$
- Relation between each value y_i and the mean $\overline{y}$

  $\longrightarrow$ Does $x_i > \overline{x}$ imply $y_i > \overline{y}$?

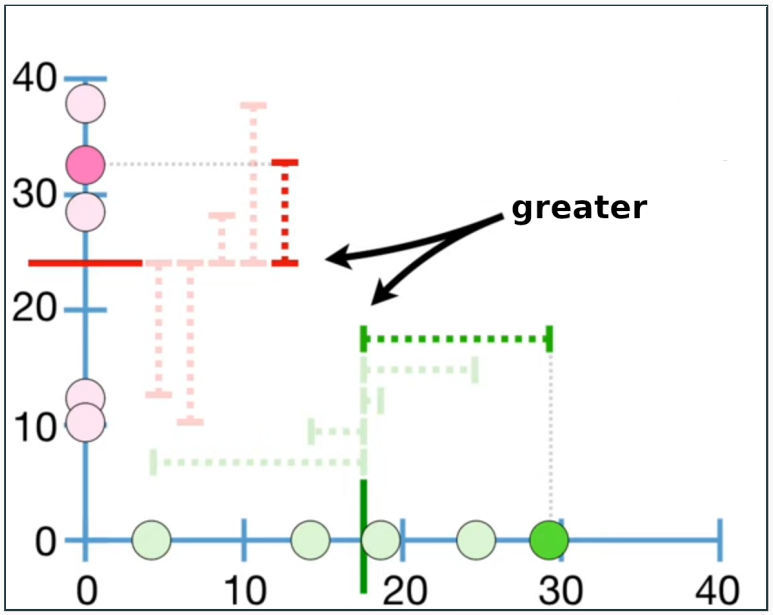
  $\longrightarrow$ Does $x_i < \overline{x}$ imply $y_i < \overline{y}$?

Covariance: far from the mean

- Relation between each value x_i and the mean $\bar{x}$
- Relation between each value y_i and the mean $\bar{y}$

  $\longrightarrow$ Does $x_i > \bar{x}$ imply $y_i > \bar{y}$?

  $\longrightarrow$ Does $x_i < \bar{x}$ imply $y_i < \bar{y}$?

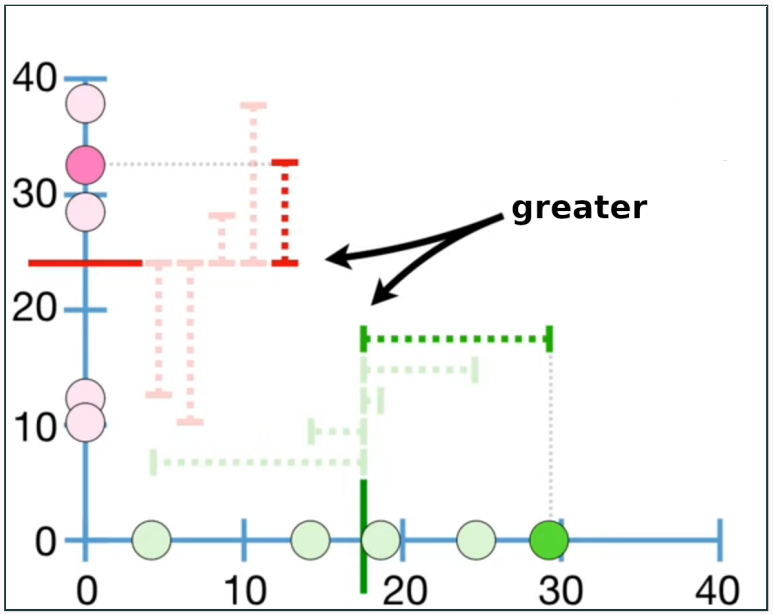


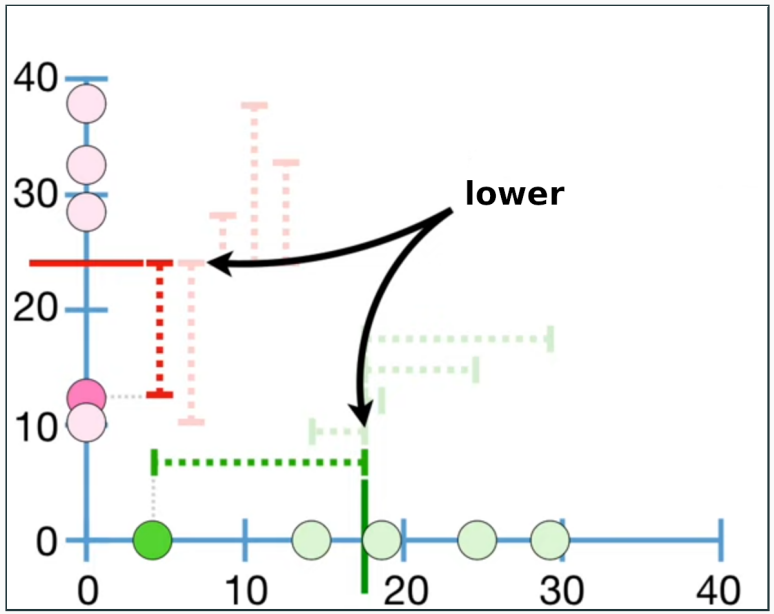
Covariance: far from the mean

- Relation between each value x_i and the mean $\bar{x}$
- Relation between each value y_i and the mean $\bar{y}$

  $\longrightarrow$ Does $x_i > \bar{x}$ imply $y_i > \bar{y}$?

  $\longrightarrow$ Does $x_i < \bar{x}$ imply $y_i < \bar{y}$?





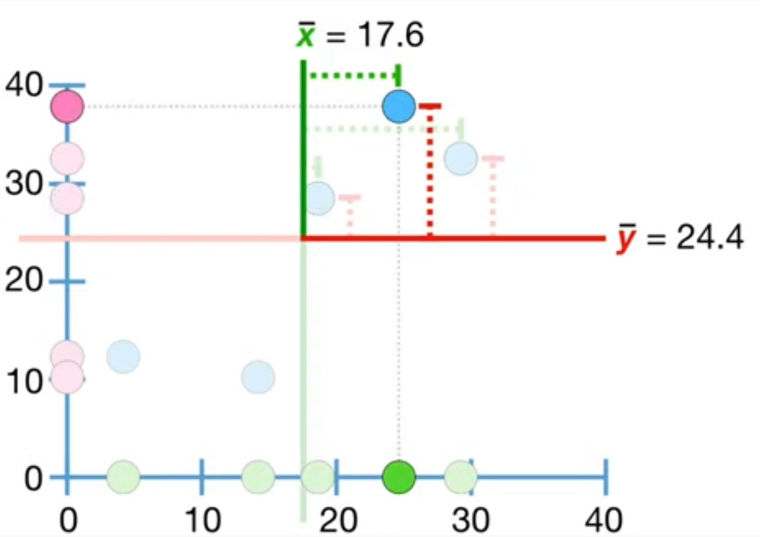
Source: https://www.youtube.com/watch?v=qtaqvPAeEJY

Covariance: vary together

- Relation between each value x_i and the mean $\overline{x}$
  - $\rightarrow x_i > \overline{x} \implies (x_i - \overline{x})$ positive
  - $\rightarrow x_i < \overline{x} \implies (x_i - \overline{x})$ negative
- Relation between each value y_i and the mean $\overline{y}$
  - $\rightarrow y_i > \overline{y} \implies (y_i - \overline{y})$ positive
  - $\rightarrow y_i < \overline{y} \implies (y_i - \overline{y})$ negative

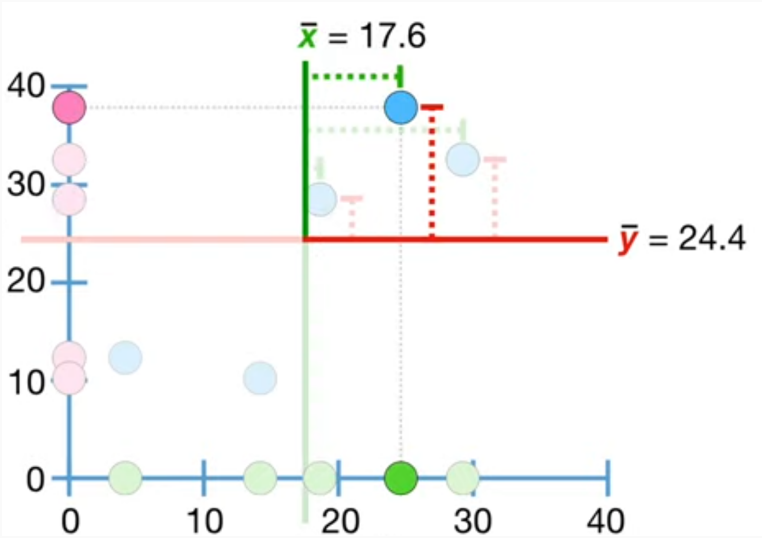
Covariance: vary together

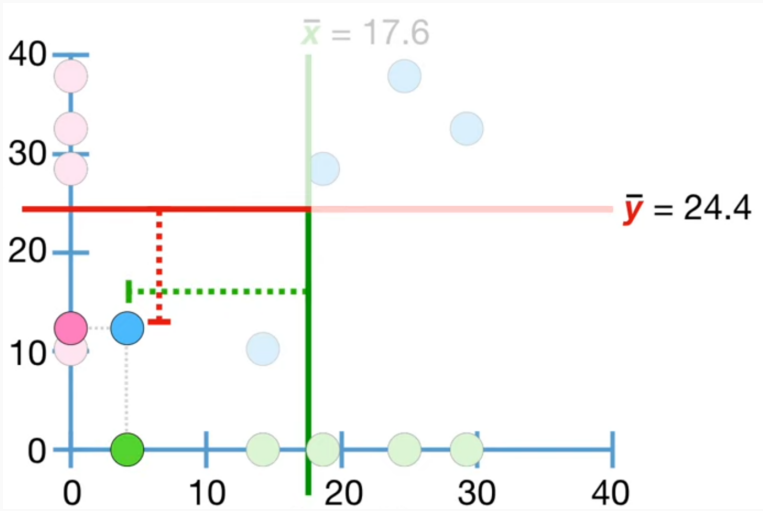
- Relation between each value x_i and the mean $\overline{x}$
  - $\rightarrow x_i > \overline{x} \implies (x_i - \overline{x})$ positive
  - $\rightarrow x_i < \overline{x} \implies (x_i - \overline{x})$ negative
- Relation between each value y_i and the mean $\overline{y}$
  - $\rightarrow y_i > \overline{y} \implies (y_i - \overline{y})$ positive
  - $\rightarrow y_i < \overline{y} \implies (y_i - \overline{y})$ negative



Covariance: vary together

- Relation between each value x_i and the mean $\bar{x}$
  - $\rightarrow x_i > \bar{x} \implies (x_i - \bar{x})$ positive
  - $\rightarrow x_i < \bar{x} \implies (x_i - \bar{x})$ negative
- Relation between each value y_i and the mean $\bar{y}$
  - $\rightarrow y_i > \bar{y} \implies (y_i - \bar{y})$ positive
  - $\rightarrow y_i < \bar{y} \implies (y_i - \bar{y})$ negative





Source: https://www.youtube.com/watch?v=qtaqvPAeEJY

Covariance: vary together

$$(x_i - \bar{x}) \times (y_i - \bar{y})$$

- Both $(x_i - \bar{x})$ and $(y_i - \bar{y})$ are positive
  $\longrightarrow$ Product $(x_i - \bar{x}) \times (y_i - \bar{y})$ is **positive**

Covariance: vary together

$$(x_i - \bar{x}) \times (y_i - \bar{y})$$

- Both $(x_i - \bar{x})$ and $(y_i - \bar{y})$ are positive
  $\longrightarrow$ Product $(x_i - \bar{x}) \times (y_i - \bar{y})$ is **positive**
- Both $(x_i - \bar{x})$ and $(y_i - \bar{y})$ are negative
  $\longrightarrow$ Product $(x_i - \bar{x}) \times (y_i - \bar{y})$ is **positive**

Covariance: vary together

$$(x_i - \bar{x}) \times (y_i - \bar{y})$$

- Both $(x_i - \bar{x})$ and $(y_i - \bar{y})$ are positive
  $\longrightarrow$ Product $(x_i - \bar{x}) \times (y_i - \bar{y})$ is **positive**
- Both $(x_i - \bar{x})$ and $(y_i - \bar{y})$ are negative
  $\longrightarrow$ Product $(x_i - \bar{x}) \times (y_i - \bar{y})$ is **positive**
- $(x_i - \bar{x})$ is positive and $(y_i - \bar{y})$ is negative
  $\longrightarrow$ Product $(x_i - \bar{x}) \times (y_i - \bar{y})$ is **negative**

Covariance: vary together

$$(x_i - \bar{x}) \times (y_i - \bar{y})$$

- Both $(x_i - \bar{x})$ and $(y_i - \bar{y})$ are positive
  $\rightarrow$ Product $(x_i - \bar{x}) \times (y_i - \bar{y})$ is **positive**
- Both $(x_i - \bar{x})$ and $(y_i - \bar{y})$ are negative
  $\rightarrow$ Product $(x_i - \bar{x}) \times (y_i - \bar{y})$ is **positive**
- $(x_i - \bar{x})$ is positive and $(y_i - \bar{y})$ is negative
  $\rightarrow$ Product $(x_i - \bar{x}) \times (y_i - \bar{y})$ is **negative**
- $(x_i - \bar{x})$ is negative and $(y_i - \bar{y})$ is positive
  $\rightarrow$ Product $(x_i - \bar{x}) \times (y_i - \bar{y})$ is **negative**

Covariance: the formula

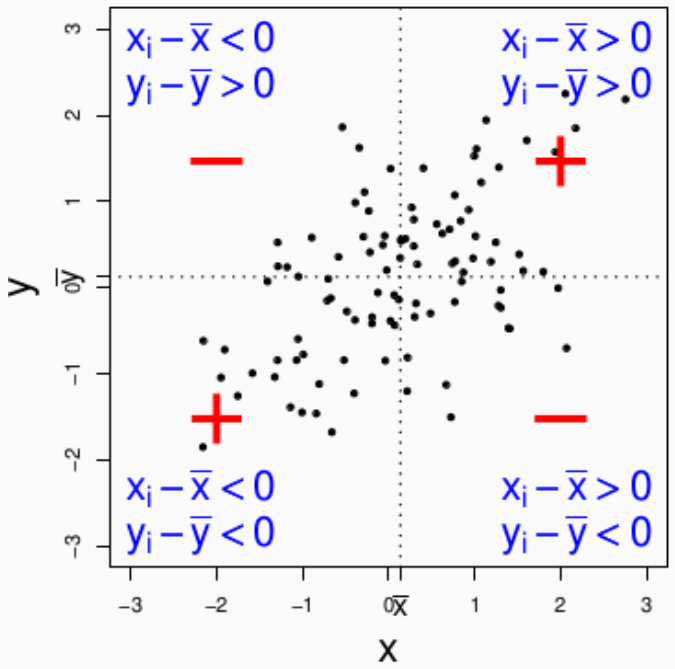
1. First calculate means $\bar{x}$ and $\bar{y}$
2. Then calculate the covariance as :

$$Cov(X, Y) = \sum_{i=1}^{n} \frac{(x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

Covariance: the formula

1. First calculate means $\bar{x}$ and $\bar{y}$
2. Then calculate the covariance as :

$$Cov(X, Y) = \sum_{i=1}^{n} \frac{(x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$



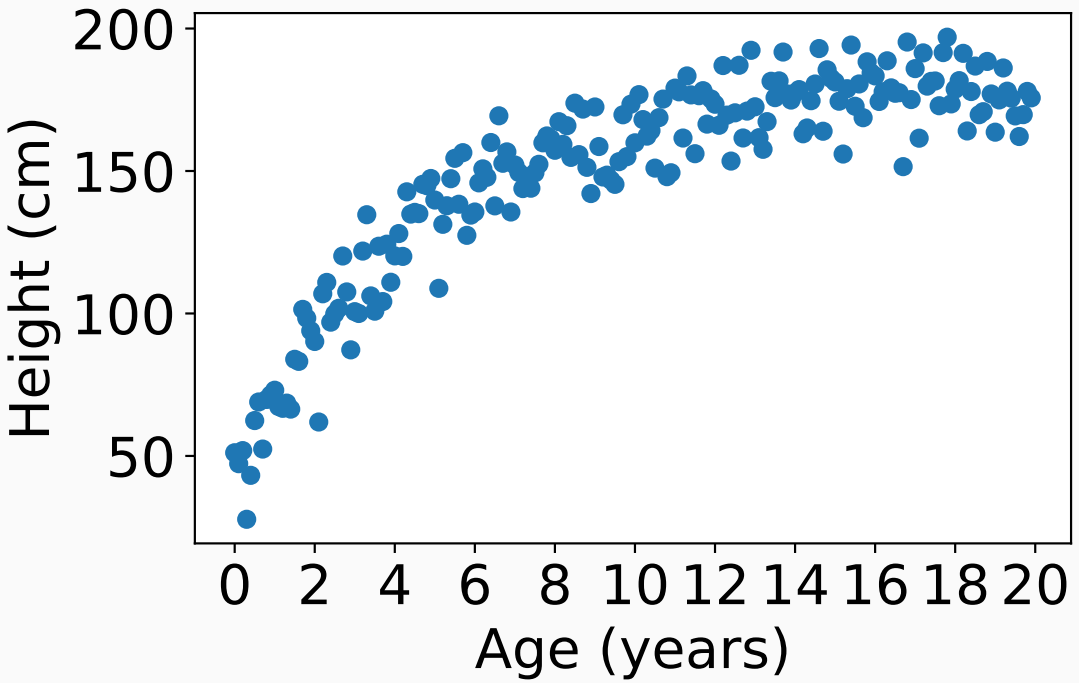
https://www.stat.uchicago.edu/~yibi/teaching/stat220/17aut/Lectures/L22.pdf

Wooclap

Wooclap time !

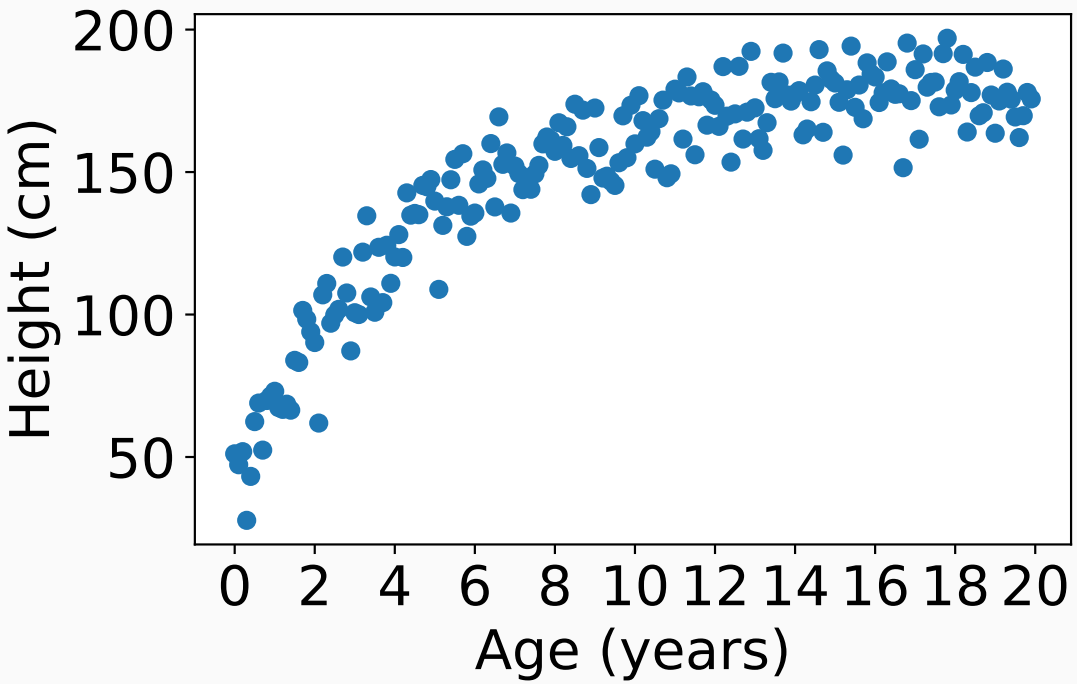
Exercise: guess the covariance

1. A person's age (X) vs. height (Y)



Exercise: guess the covariance

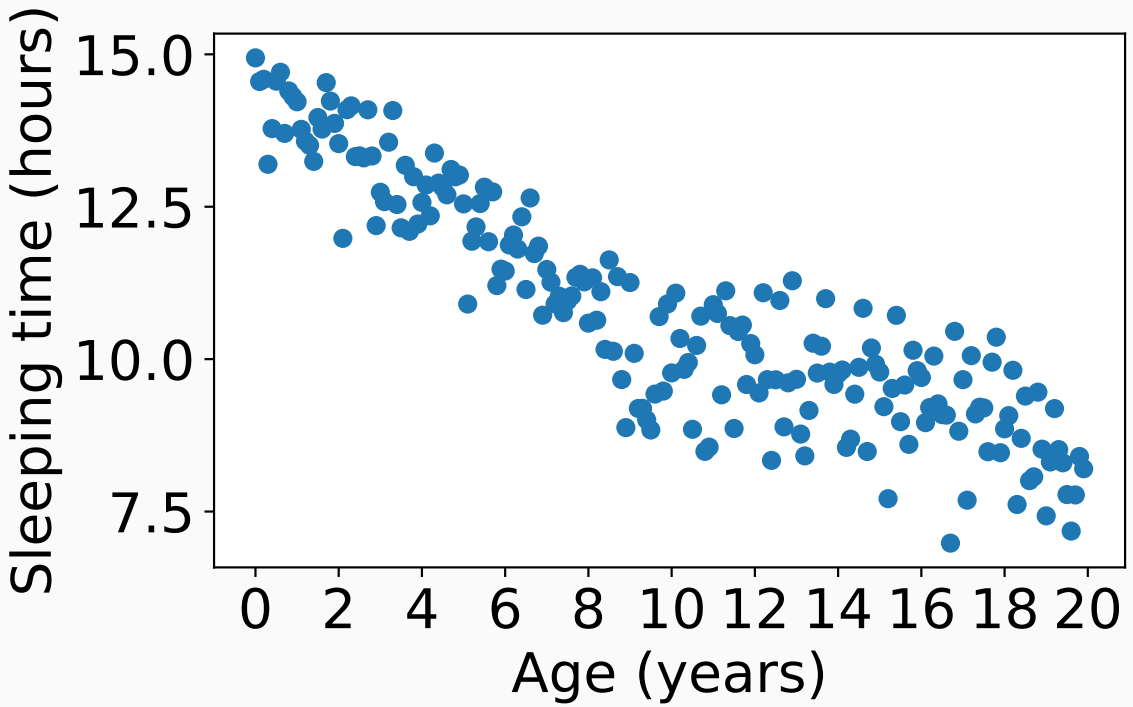
1. A person's age (X) vs. height (Y)



$$Cov(X, Y) = +180.9$$

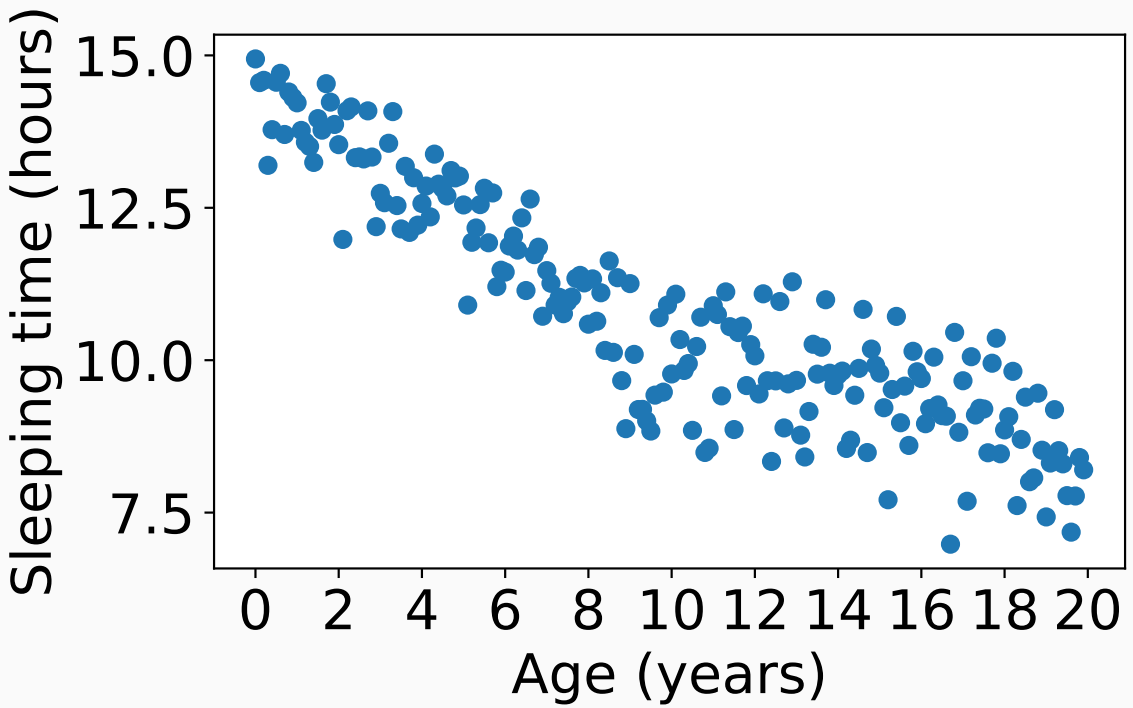
Exercise: guess the covariance

A person's age (X) vs. number of sleeping hours (Y)



Exercise: guess the covariance

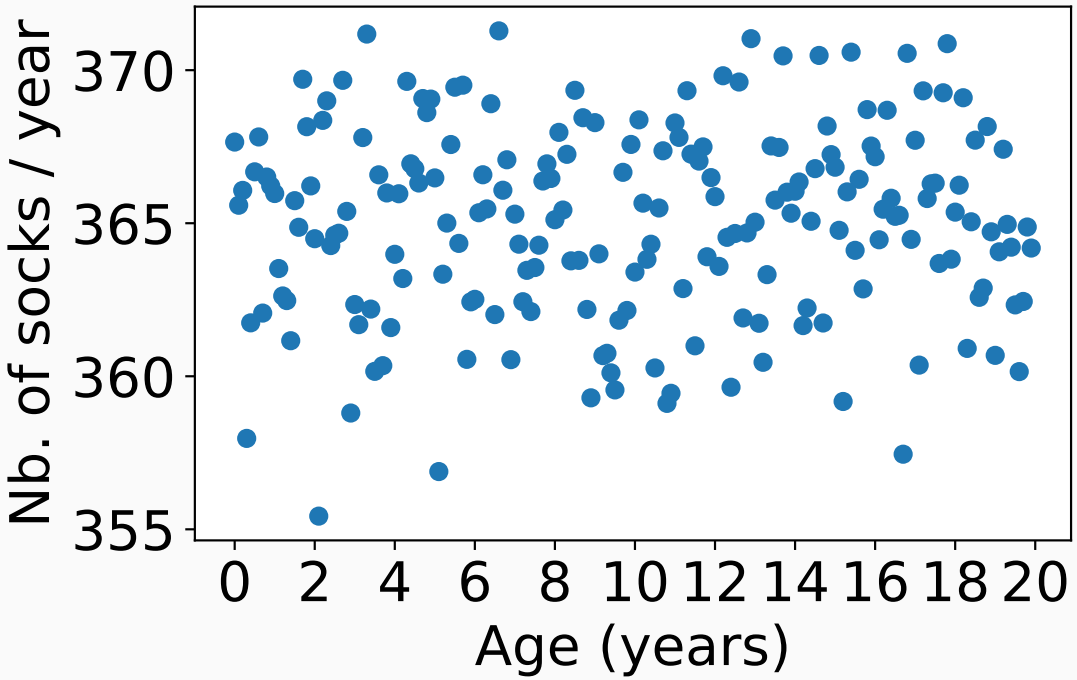
A person's age (X) vs. number of sleeping hours (Y)



$$Cov(X, Y) = -9.0$$

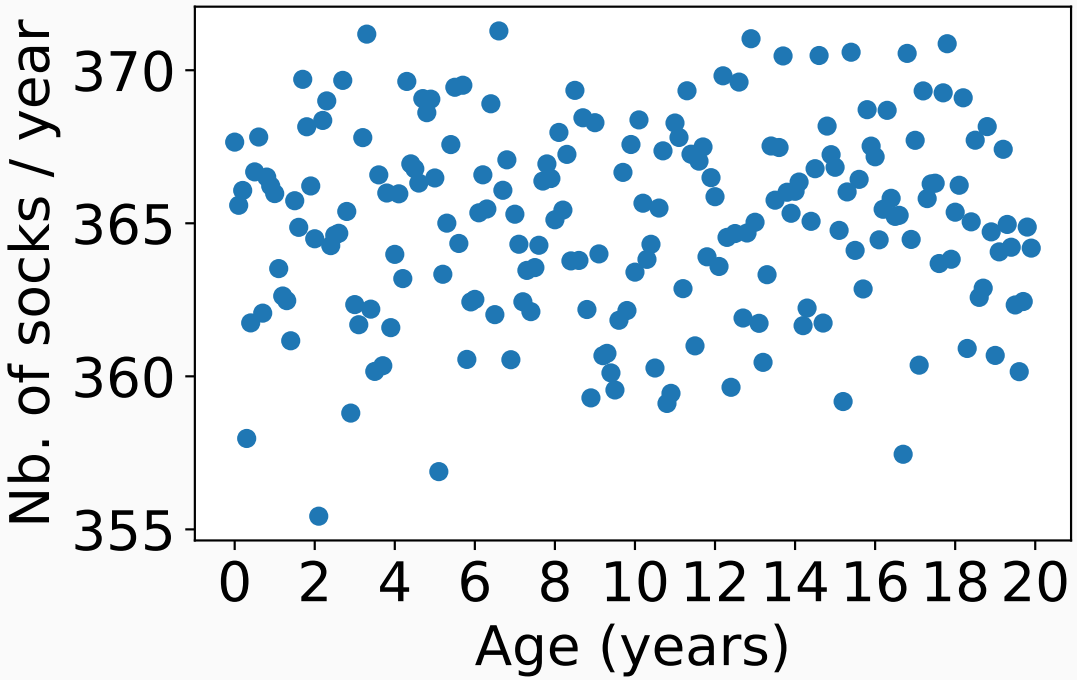
Exercise: guess the covariance

A person's age (X) vs. number of socks used per year (Y)



Exercise: guess the covariance

A person's age (X) vs. number of socks used per year (Y)



$$Cov(X, Y) = 0.77$$

Covariance is sensitive to unit

- What if X and Y have very different ranges?

  $\rightarrow$ For instance, X in cm, Y in km

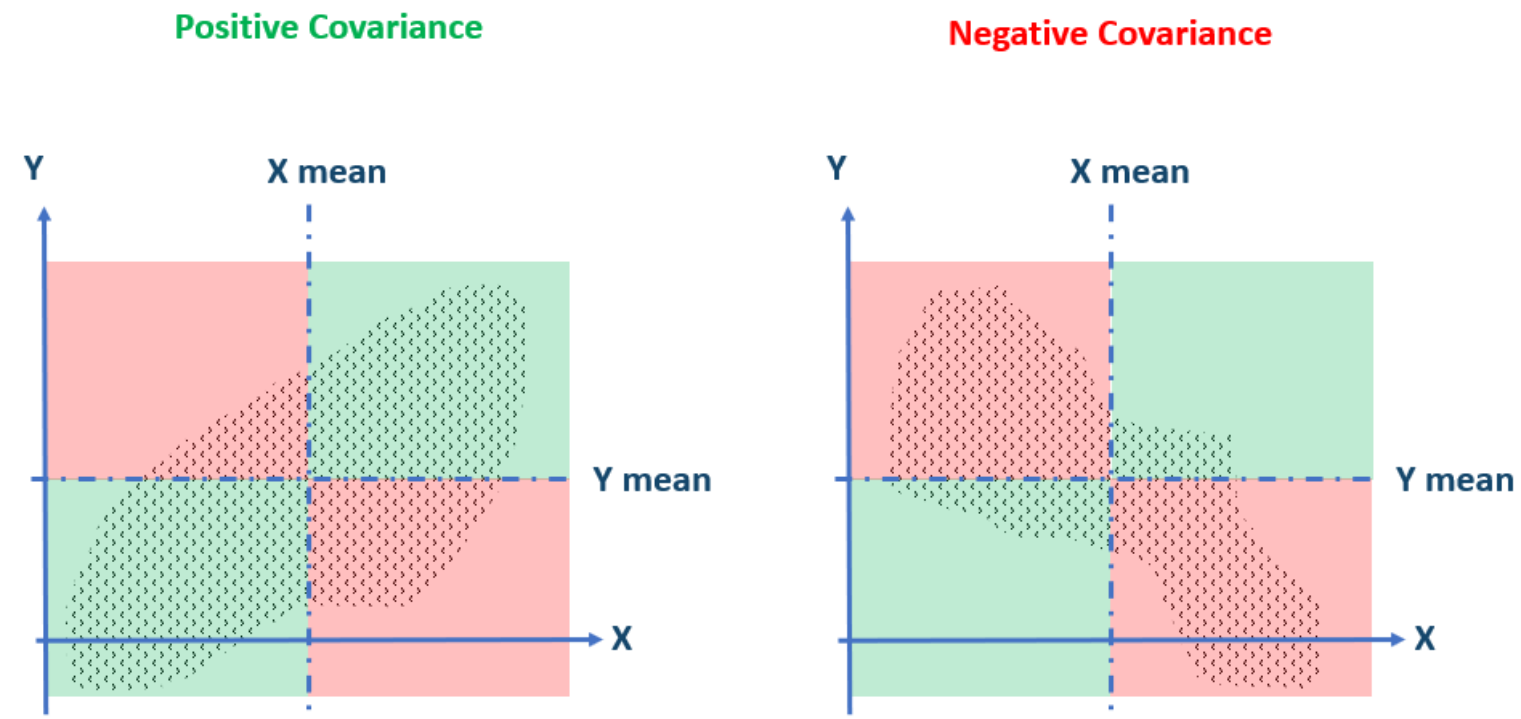
Covariance is sensitive to unit

- What if X and Y have very different ranges?

  $\rightarrow$ For instance, X in cm, Y in km

- Covariance is unbounded - ranges from $-\infty$ to $+\infty$

  $\rightarrow$ Indicates whether a linear relation exists, but not its strength



Covariance: it's a sign!

- Covariance is **positive**
  - $\rightarrow$ Increasing X tends to make Y increase too
- Covariance is **negative**
  - $\rightarrow$ Increasing X tends to make Y decrease
- Covariance is **zero**
  - $\rightarrow$ Increasing X has no impact on Y
  - $\rightarrow$ Increasing Y has no impact on X

What if. . .

- What if we could normalise covariance?
- Can we get a measure that is bounded?

Correlation coefficient (r)

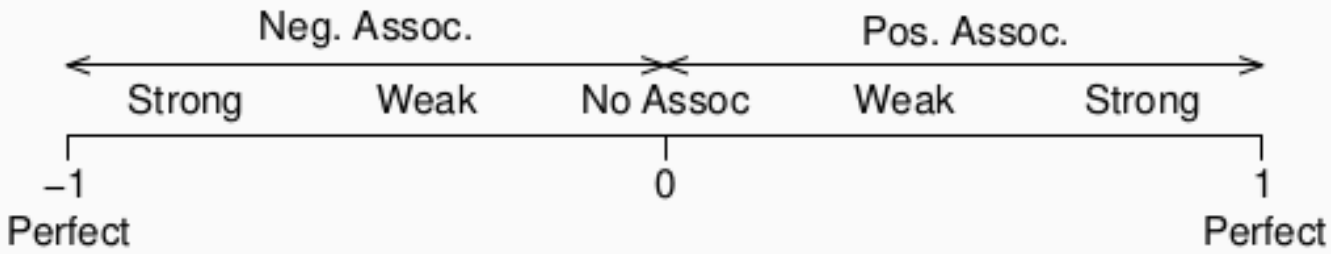
- Covariance can be normalised using X and Y's **variances**

$$r_{X,Y} = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y))}} = \frac{Cov(X,Y)}{s_X s_Y}$$

- Dividing by standard deviation puts both on same scale
- Also called Pearson or linear correlation

Correlation interpretation

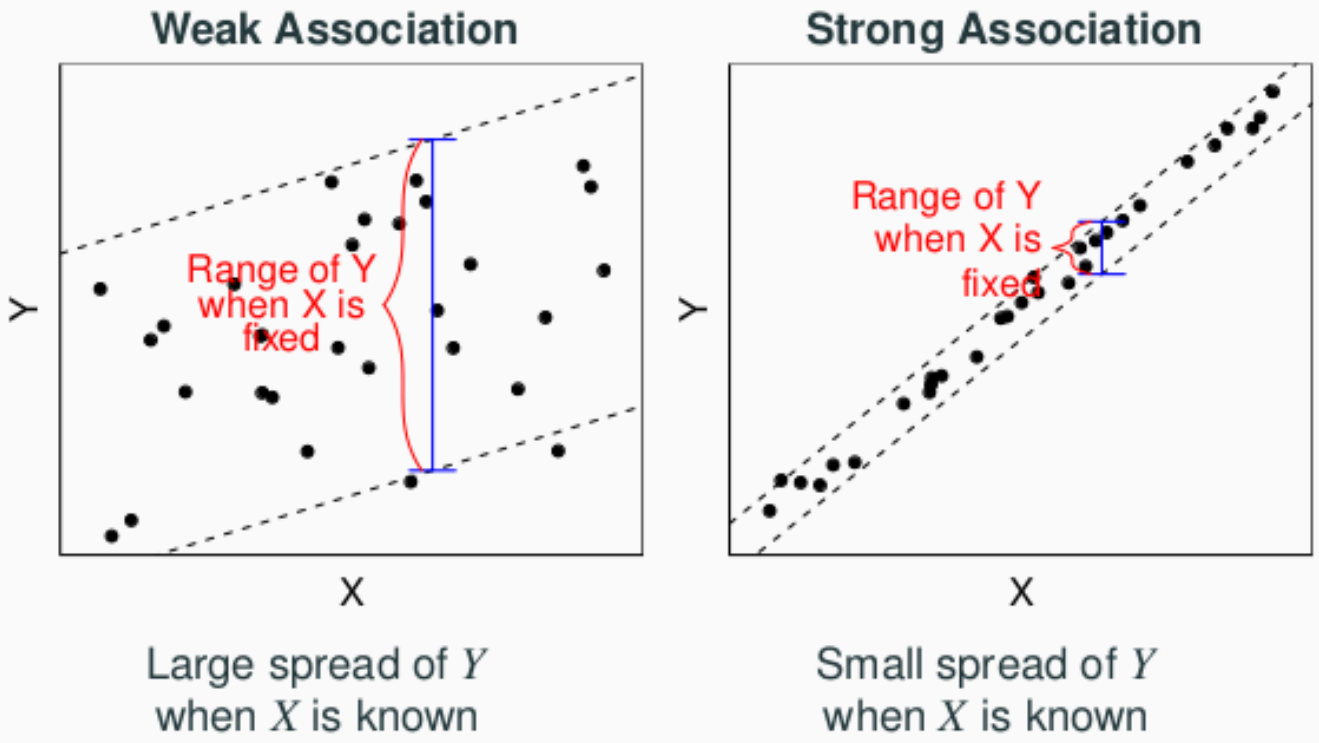
- Ranges from -1 to $+1$
  - $\rightarrow$ $r \approx +1$: strong positive association
  - $\rightarrow$ $r \approx -1$: strong negative association
  - $\rightarrow$ $r \approx 0$: weak/no linear relationship



https://www.stat.uchicago.edu/~yibi/teaching/stat220/17aut/Lectures/L22.pdf

Correlation and spread

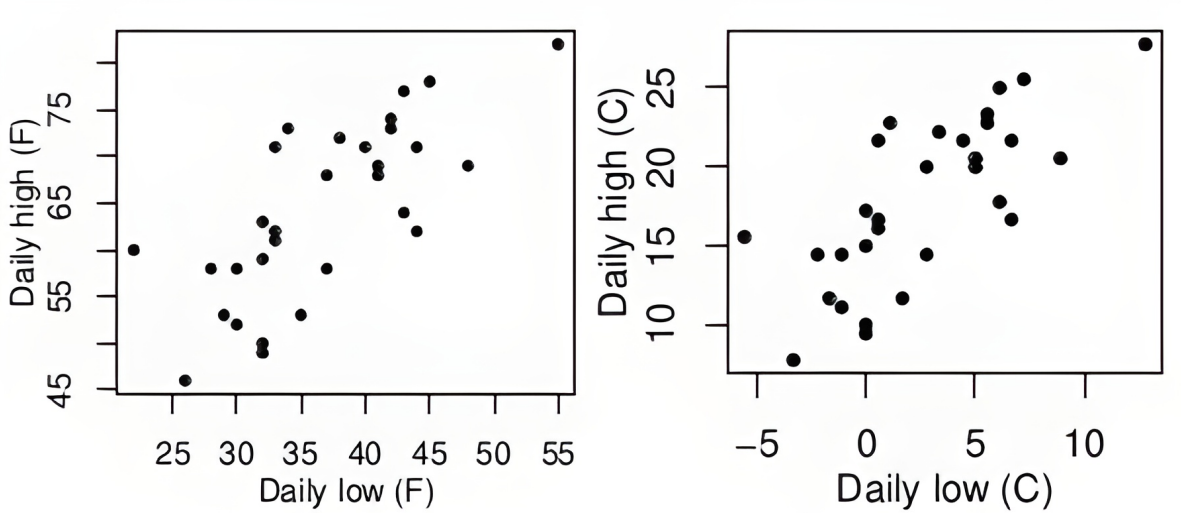
- Correlation tells how close or far from linear regression line
  - $\longrightarrow$ Knowing x allows predicting y (and vice-versa)
  - $\longrightarrow$ Coefficient of determination R^2 = square of Pearson correlation r



Source: https://www.stat.uchicago.edu/~yibi/teaching/stat220/17aut/Lectures/L22.pdf

Correlation is unit-less

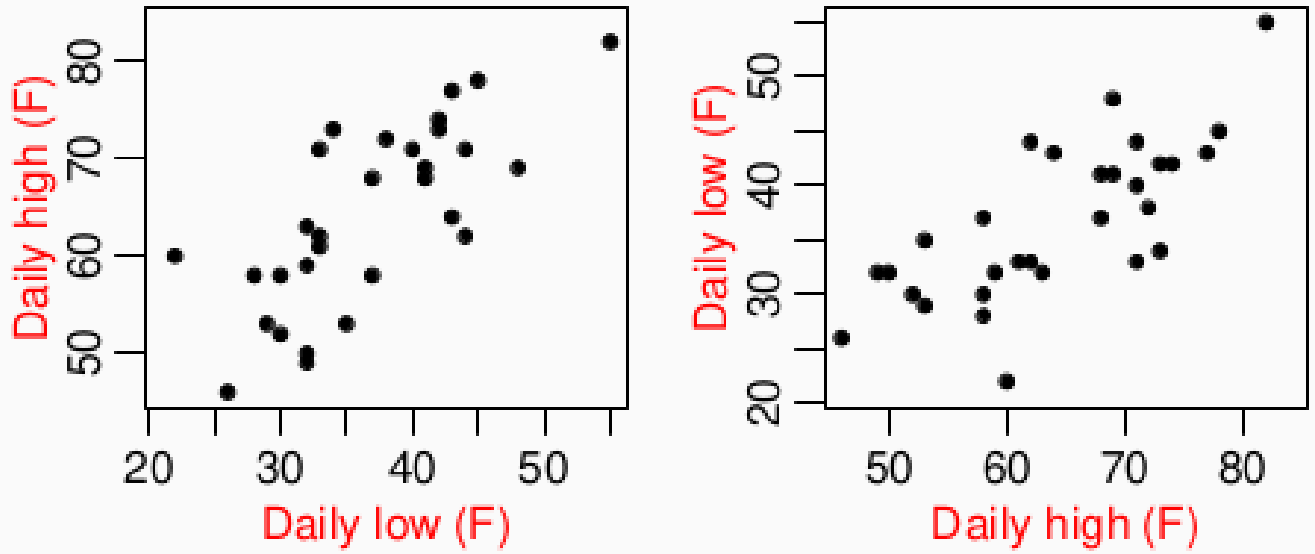
- Covariance is unbounded, depends on variable ranges
- Correlation allows comparing metrics with different ranges
  - $\rightarrow$ Example: max vs. min. temperature in Celsius or Farehnheit
  - $\rightarrow$ In both cases, correlation is the same: $r = 0.74$



https://www.stat.uchicago.edu/~yibi/teaching/stat220/17aut/Lectures/L22.pdf

Correlation is symmetric

- Correlation is symmetric
  - → Example: max vs. min. temperature or vice-versa
  - → In both cases, correlation is the same: $r = 0.74$



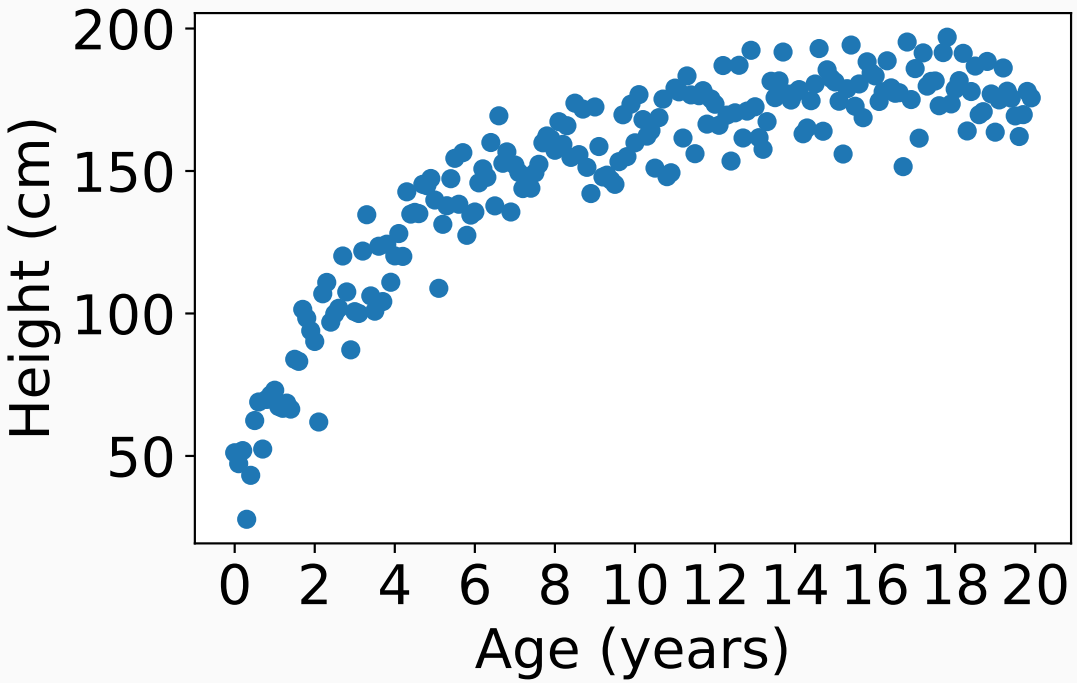
https://www.stat.uchicago.edu/~yibi/teaching/stat220/17aut/Lectures/L22.pdf

Wooclap

Wooclap time !

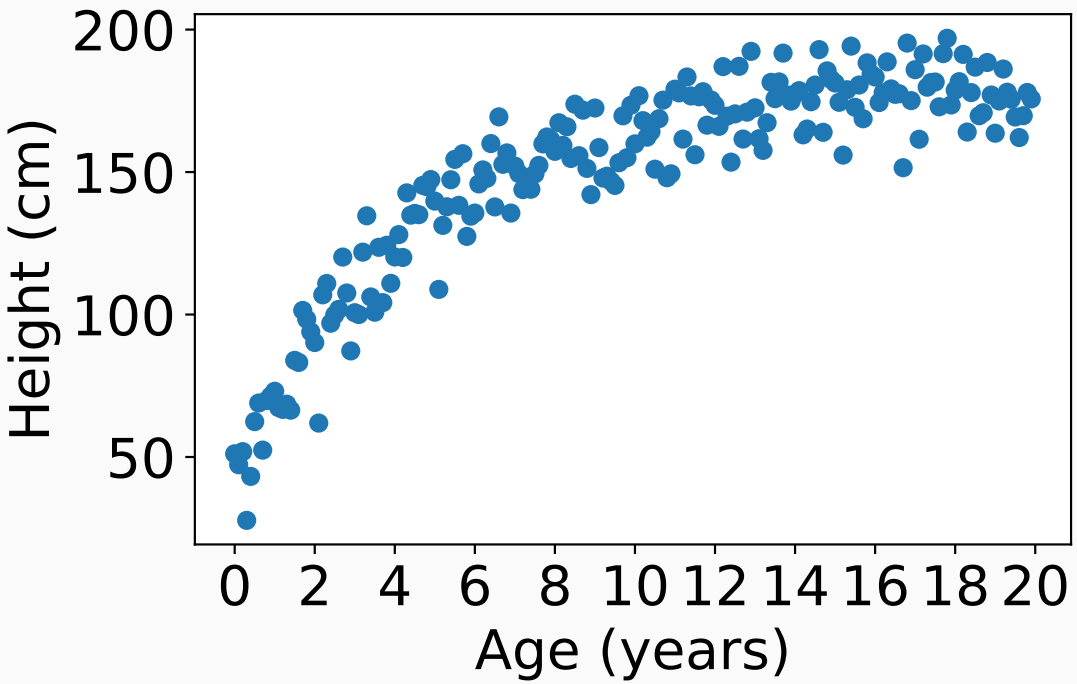
Exercise: guess the correlation

1. A person's age (X) vs. height (Y)



Exercise: guess the correlation

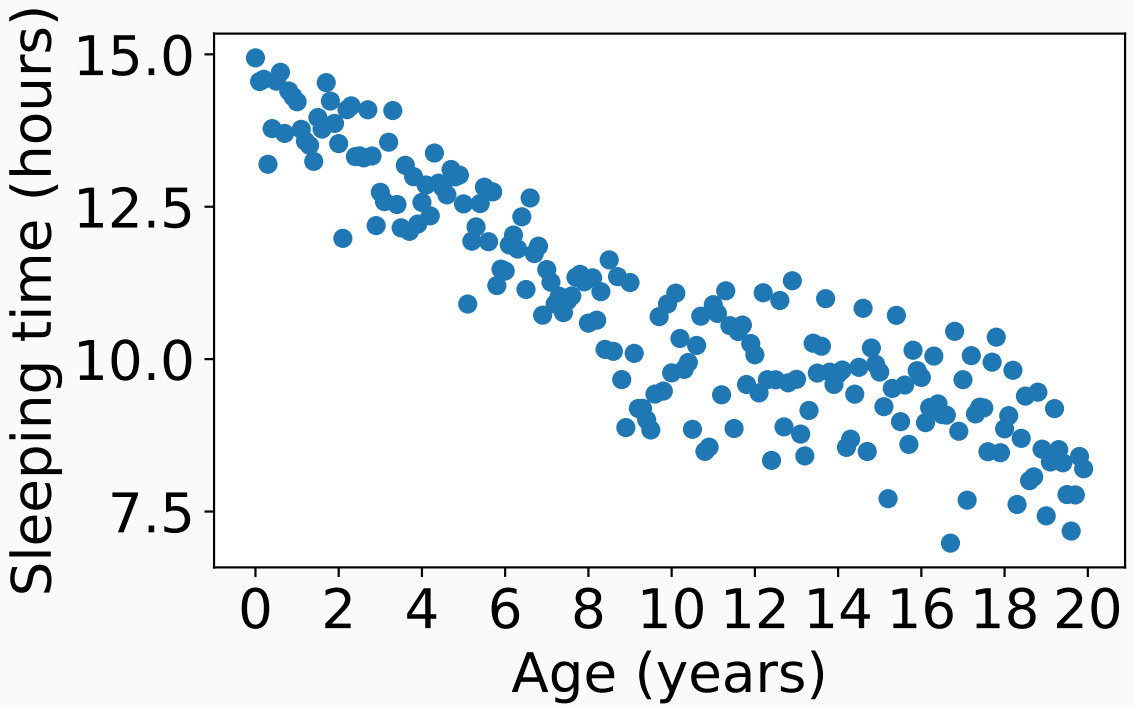
1. A person's age (X) vs. height (Y)



$$r(X, Y) = 0.85$$

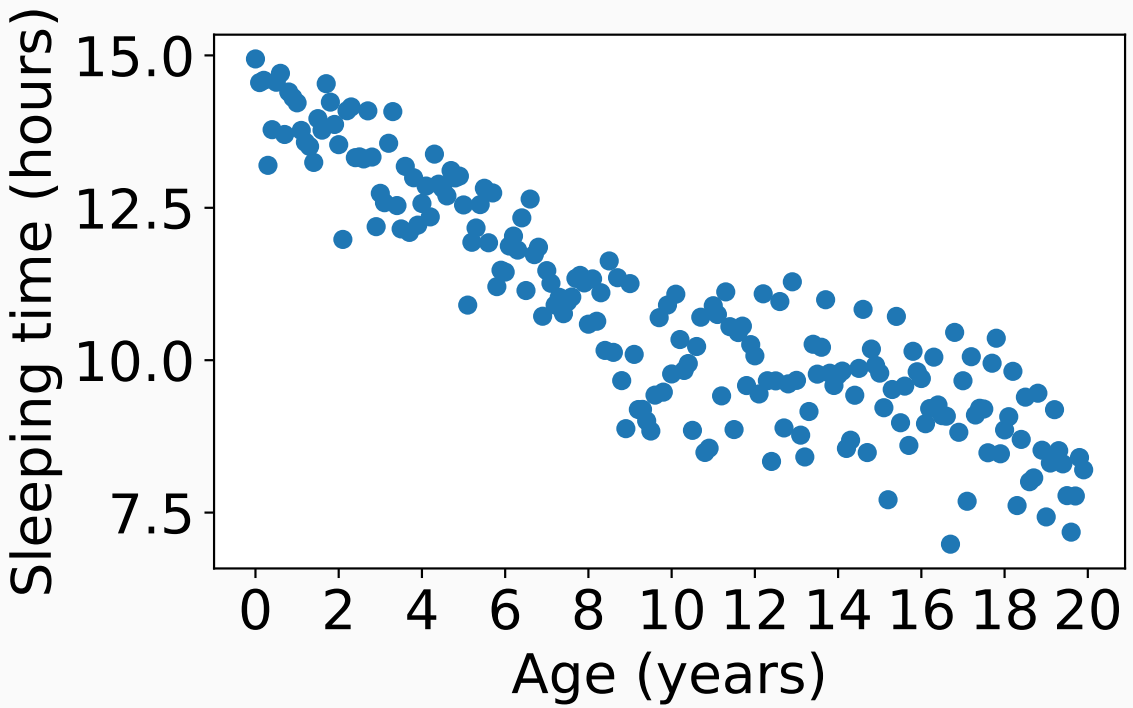
Exercise: guess the correlation

A person's age (X) vs. number of sleeping hours (Y)



Exercise: guess the correlation

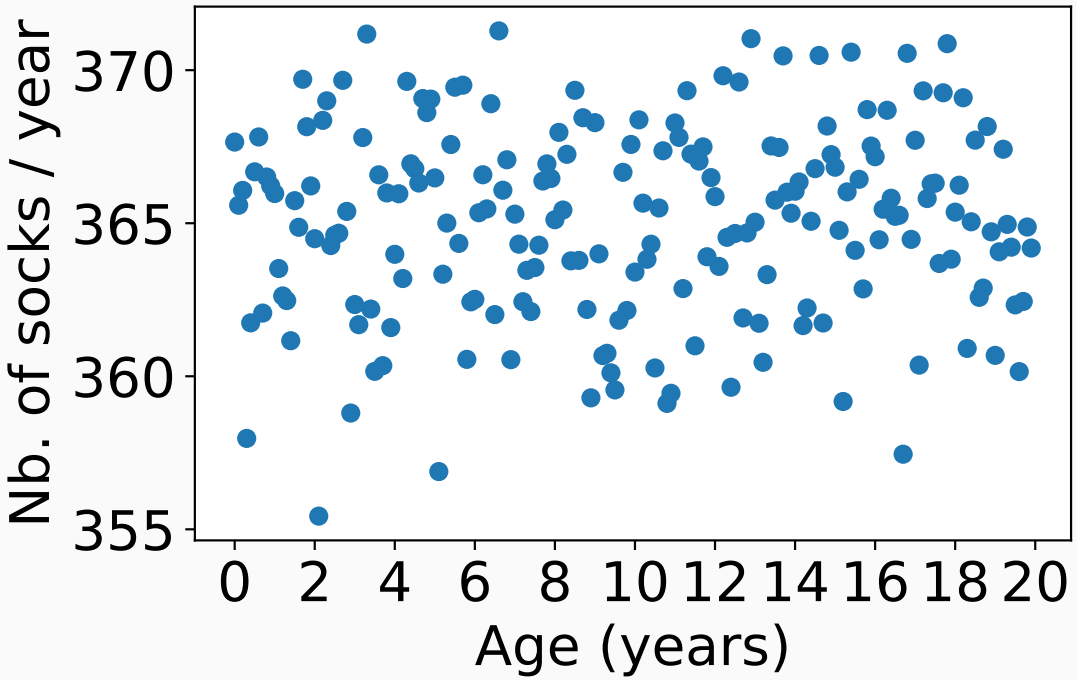
A person's age (X) vs. number of sleeping hours (Y)



$$r(X, Y) = -0.89$$

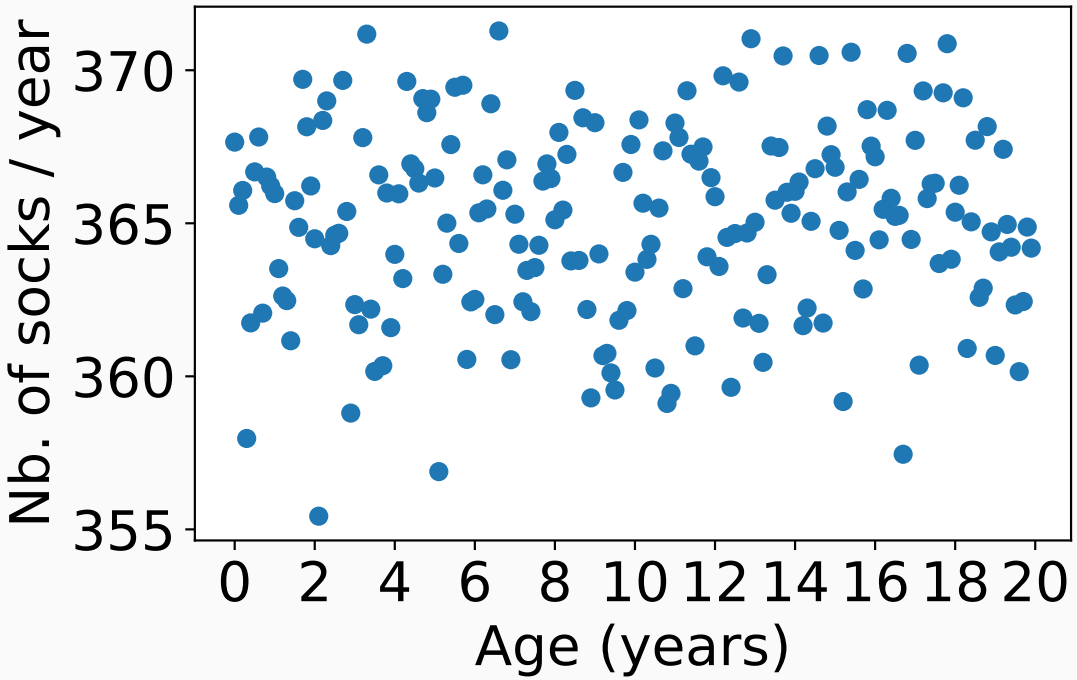
Exercise: guess the correlation

A person's age (X) vs. number of socks used per year (Y)



Exercise: guess the correlation

A person's age (X) vs. number of socks used per year (Y)

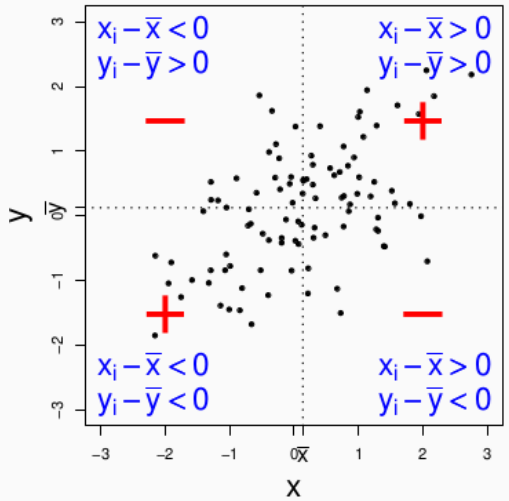


$$r(X, Y) = 0.04$$

Why dividing by standard deviations?

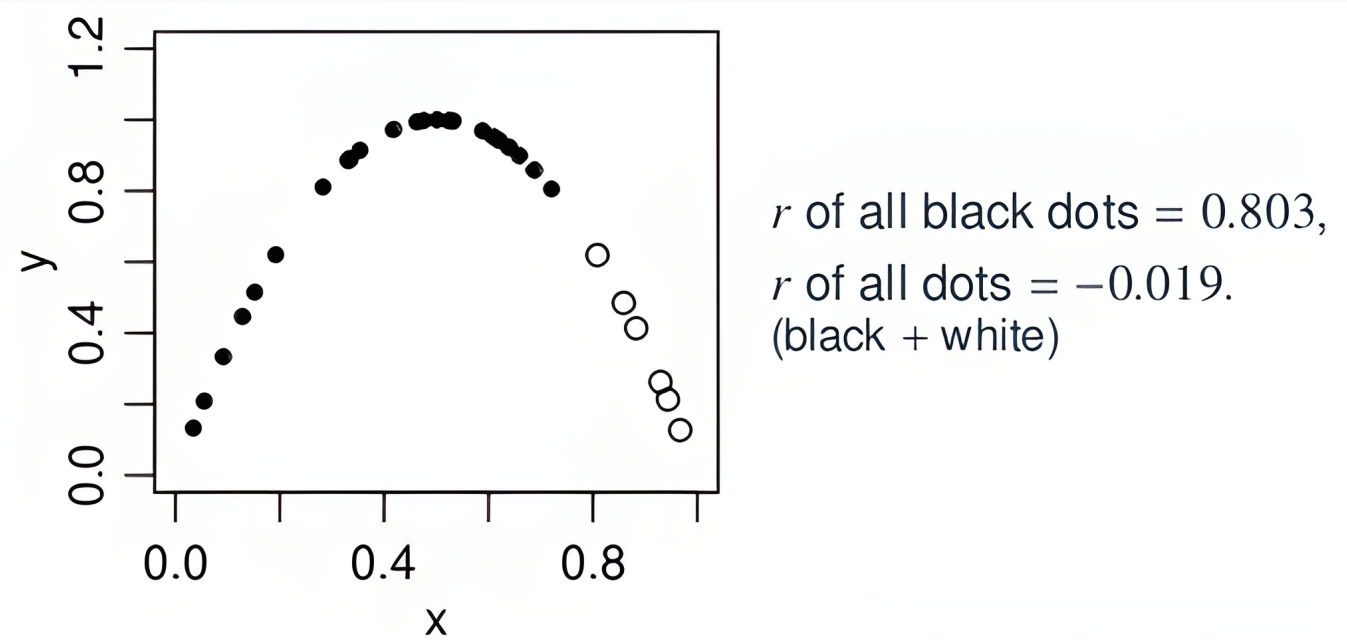
$$r_{X,Y} = \frac{Cov(X,Y)}{s_X s_Y} = \frac{1}{n-1}\frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{s_X s_Y}$$
$$= \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}}{s_X}\right)\left(\frac{y_i - \overline{y}}{s_Y}\right)$$

- Similar to standardisation in normal distribution
  - $\longrightarrow$ Discounting the mean centers around zero
  - $\longrightarrow$ Dividing by standard deviation homogenizes width



Correlation shows linear association

- Correlation does not model non-linear association



https://www.stat.uchicago.edu/~yibi/teaching/stat220/17aut/Lectures/L22.pdf

Correlation of compositionality

Jupyter notebook 8

- Hypothesis: compositionality and frequency are correlated
  - $\rightarrow$ Frequency is better represented in logarithmic scale
- Does correlation change if frequency is in linear or log scale?

Spearman's rank correlation

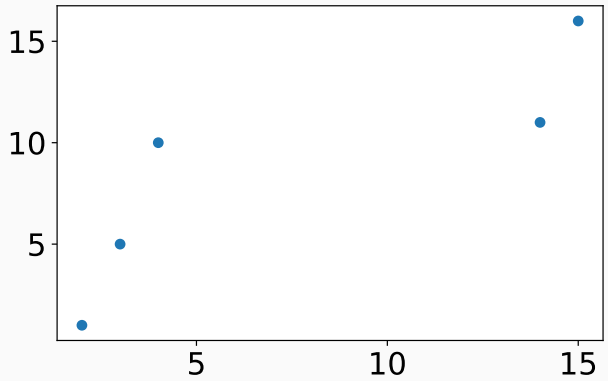
- The actual compared X and Y values may be irrelevant
  - $\rightarrow$ Does X rank items more or less in the same order as Y?
- Spearman's ρ: linear (Pearson) correlation between ranks
  - $\rightarrow$ Models monotonic relation

Spearman's rank correlation

- The actual compared X and Y values may be irrelevant
  - $\rightarrow$ Does X rank items more or less in the same order as Y?
- Spearman's ρ: linear (Pearson) correlation between ranks
  - $\rightarrow$ Models monotonic relation

Example:

```
x = [2,3,4,14,15]
y = [1,5,10,11,16]
```



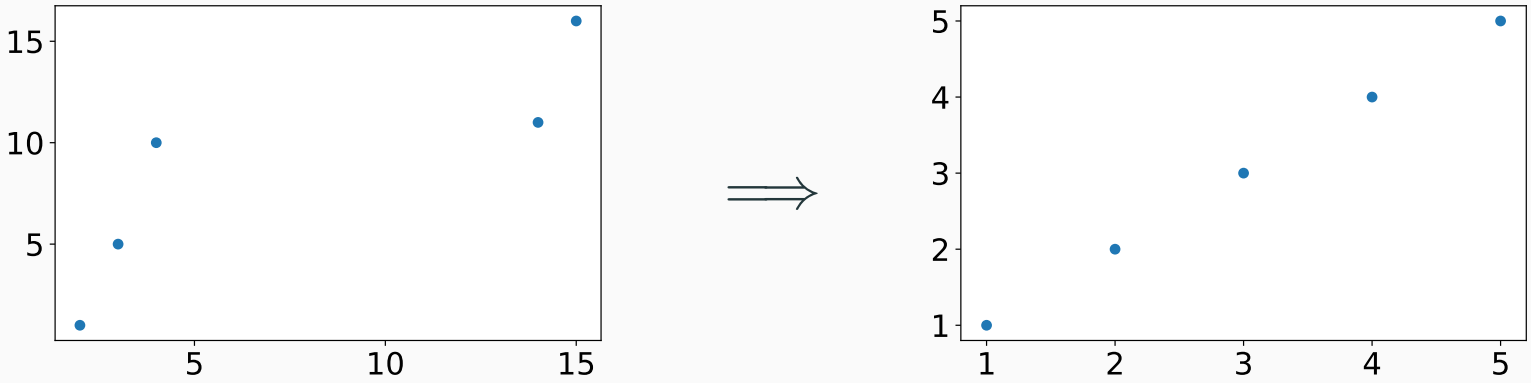
$\Longrightarrow$

Spearman's rank correlation

- The actual compared X and Y values may be irrelevant
  - $\rightarrow$ Does X rank items more or less in the same order as Y?
- Spearman's ρ: linear (Pearson) correlation between ranks
  - $\rightarrow$ Models monotonic relation

Example:

```
x = [2,3,4,14,15]
y = [1,5,10,11,16]
```



Spearman correlation

- Obtain ranks rX_i for X in ascending order
- Obtain ranks rY_i for Y in ascending order
- Obtain difference between ranks $d_i = rX_i - rY_i$
- Calculate Spearman's rank correlation:

$$\rho_{X,Y} = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

Spearman correlation

- Obtain ranks rX_i for X in ascending order
- Obtain ranks rY_i for Y in ascending order
- Obtain difference between ranks $d_i = rX_i - rY_i$
- Calculate Spearman's rank correlation:

$$\rho_{X,Y} = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

- Alternatively, Pearson correlation between rX_i and rY_i

Spearman correlation: example

IQ, X_i	Hours of TV per week, Y_i	rank x_i	rank y_i	d_i	d_i^2
86	2	1	1	0	0
97	20	2	6	−4	16
99	28	3	8	−5	25
100	27	4	7	−3	9
101	50	5	10	−5	25
103	29	6	9	−3	9
106	7	7	3	4	16
110	17	8	5	3	9
112	6	9	2	7	49
113	12	10	4	6	36

Source: `https://en.wikipedia.org/wiki/Spearman_correlation`

Pearson vs. Spearman of compositionality

Jupyter notebook 9 & 10

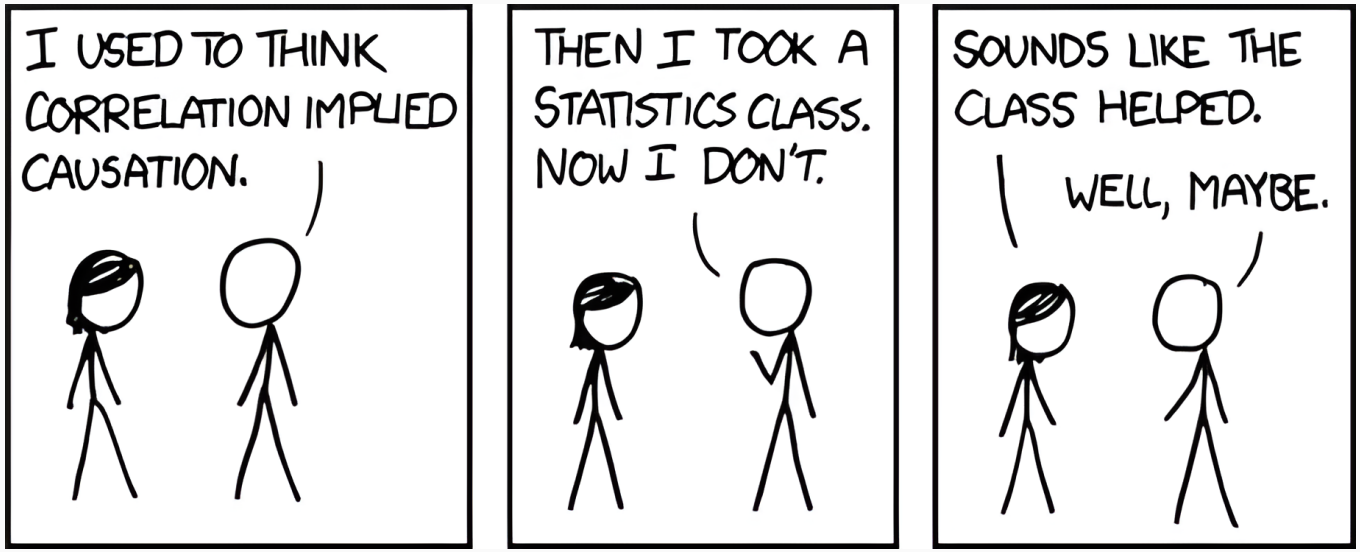
- Compare Pearson and Spearman correlation
  - $\longrightarrow$ Compositionality vs. frequency
  - $\longrightarrow$ Compositionality vs. log-frequency
- Compare manual implementation and scipy

Warning!

Confounders

- Suppose X independent and Y dependent variables
- A confounder can influence both X and Y

Correlation is not causation!

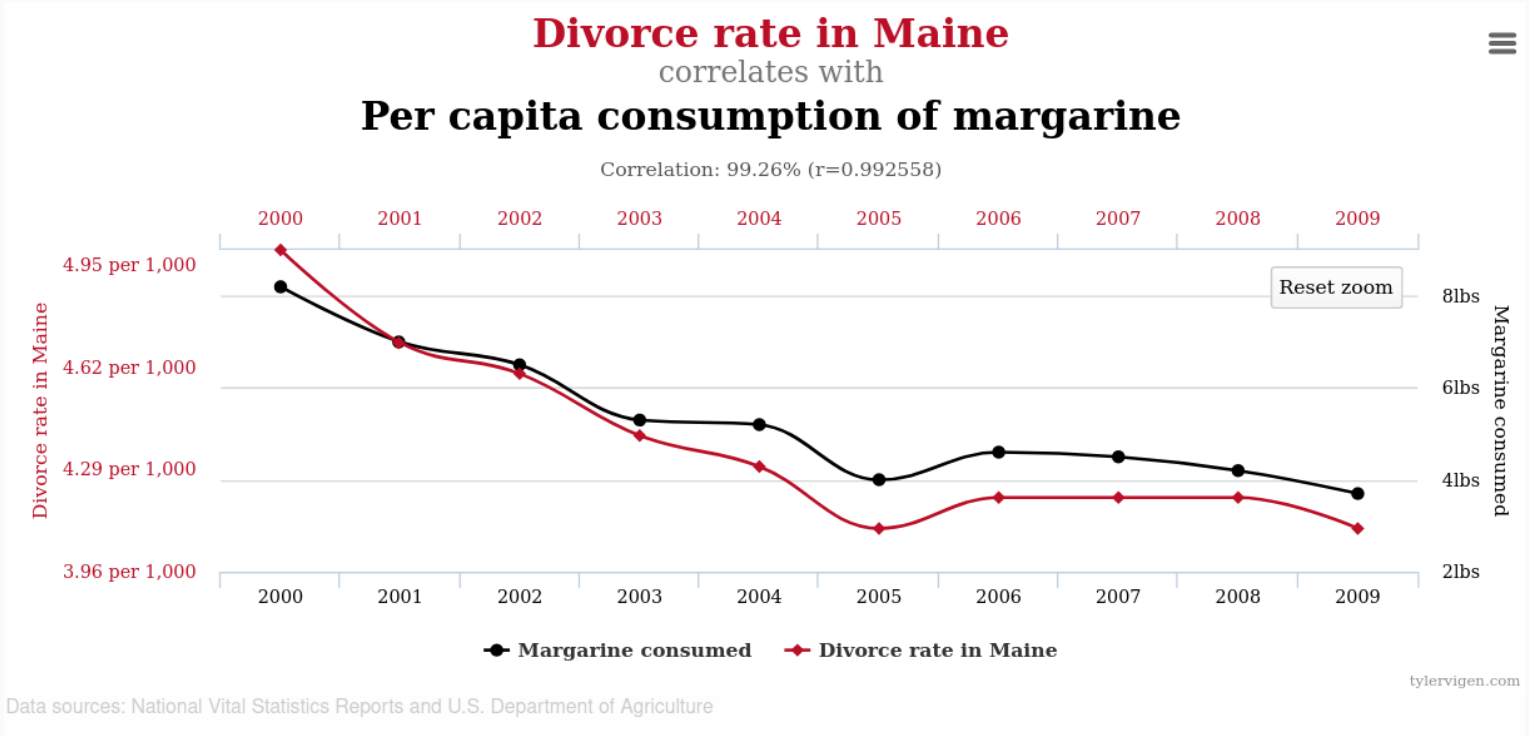


Source: https://xkcd.com/552/

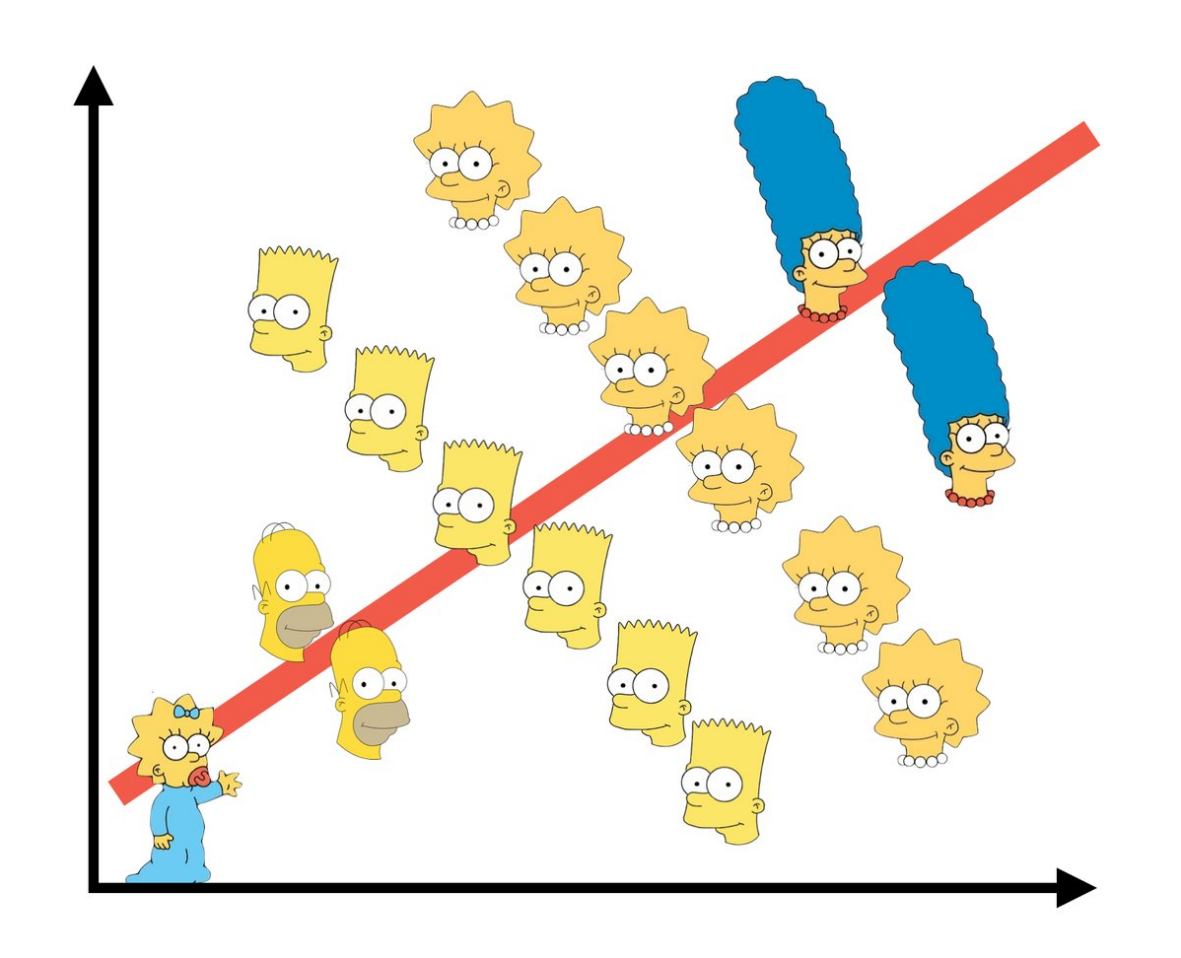
Spurious correlations

- Correlations can be found between unrelated variables
- Procrastinate: `https://www.tylervigen.com/spurious-correlations`

  → What possible confounders could explain these correlations?



Simpson's paradox



Suggestion: `https://www.arte.tv/fr/videos/107398-002-A/voyages-au-pays-des-maths/`

Outline

From raw numbers to conclusions

- Raw experimental results are usually awful to look at
- Analysis: from results (numbers) to conclusions (sentences)
  - → Choosing the most interesting and relevant results
  - → Using statistical tools such as correlation, significance, etc.
  - → Using visualisations such as tables and charts

Data analysis: hands-on activity

- Goal: apply statistical tools to real data
- How: A large table with results is provided
  - → 1. Try to identify correlations and significant differences
  - → 2. Describe what can be concluded from these statistics
  - → 3. Identify what you cannot conclude, and why (e.g. missing data)
- You should work in groups (max. 2 people) or alone
- Write a report (2-4 pages) describing analyses and conclusions

Example:

> *Model A obtains an average accuracy of 56.7 over all 14 languages, whereas model B obtains 55.9. A two-tailed t-test for paired samples gives a p-value of 0.003, below the 0.05 significance threshold. Thus, we can conclude that model A is . . .*

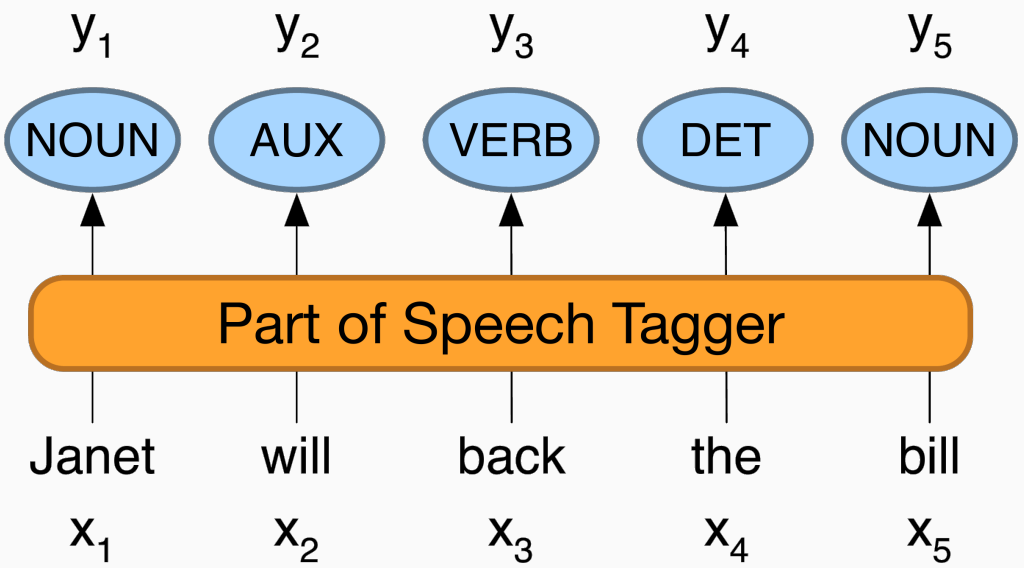
Raw data table

https://docs.google.com/spreadsheets/d/1EWWOo77PCmKfD_amwoEeUm53u_4Og3omIf5VwKFcxiM/edit?usp=sharing

Source: Manon Scholivet's thesis (all data + most descriptions, schemas, slides)

Data description

- Results of POS tagging for 38 languages
  - $\rightarrow$ POS tag = word category (noun, verb. . .)
- Each language was evaluated on a standardised test set
- Column pair: accuracy and standard deviation per language



Source: Jurafsky & Martin (2024)

Evaluation metric

Accuracy: proportion of correct predictions

	Le	poulpe	mange	le	crabe	.
Ref.	DET	NOUN	VERB	DET	NOUN	PUNCT
Pred.	VERB	NOUN	VERB	NOUN	NOUN	PUNCT

$$\frac{nb\_correct\_tags \times 100}{nb\_tags} = \frac{4 \times 100}{6} = 66.67\%$$

Source: Manon Scholivet's thesis slides

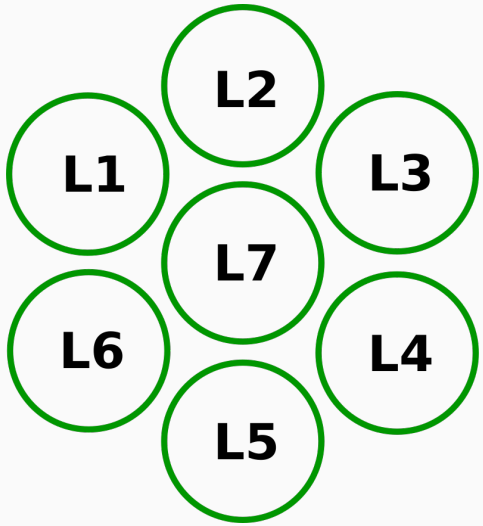
Experimental conditions i

Mono setting

- Monolingual: train on data in the test set language

  $\longrightarrow$ 38 different models, one per row

- E.g. train only on Arabic, test on... Arabic!



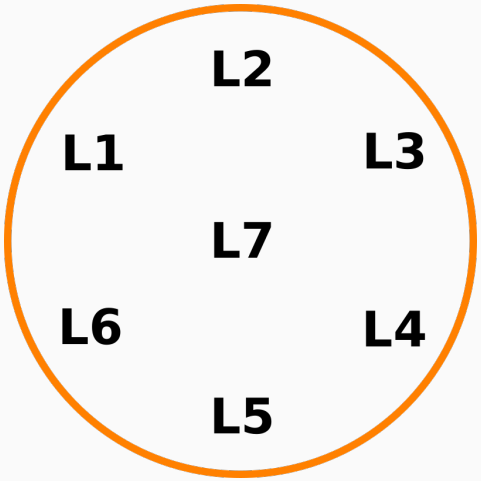
Experimental conditions ii

Multi setting

- Train on all 38 languages at once, test on single language

  $\rightarrow$ A single model trained on large multilingual corpus

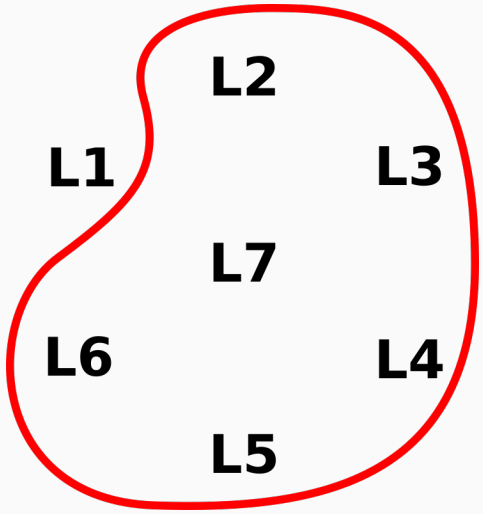
- E.g. train on 38 languages (including Arabic), test on Arabic



Experimental conditions iii

ZS setting

- Zero-shot: simulate absence of training data in the language
  - $\rightarrow$ Train on all 37 languages except target
  - $\rightarrow$ 38 different models, one per row
- E.g. train on all languages except Arabic, test on Arabic!



Experimental conditions iv

$+c$ and $-c$ settings

- Presence ($+c$) and absence ($-c$) of character information
  - $\rightarrow$ Each word is represented by the characters composing it
  - $\rightarrow$ Similar languages with same/different alphabets?

ID setting

- Presence of a number to identify the language of each word

W_{22} setting

- Presence of a typological vector for each word
  - $\rightarrow$ 22 yes/no values from World Atlas of Language Structures
  - $\rightarrow$ E.g. word from subject-verb or verb-subject language?

Suggestions

- Download the spreadsheet and use your favorite software
  - → Excel, Libreoffice, pandas, jupyter notebook, R...
- If you use Python, here are a few useful functions:
  - → `df=pandas.read_csv('filename.tsv', sep='\t')`
  - → `scipy.stats.ttest_rel()`: paired t-test
  - → `scipy.stats.pearsonr()`: Pearson correlation
  - → `scipy.stats.spearmanr()`: Spearman correlation
- Comparing `*_acc` columns is probably most interesting
- Relate your findings to explicit research (sub-)questions
- Make choices and justify them concisely
- You can use tables and charts to illustrate your report

Thanks!

That's all for today

Carlos Ramisch
first.last@lis-lab.fr

M2 IAAA - based on the course *Zen Research*
By Carlos Ramisch and Manon Scholivet

Sources i

- Adeline Paiement's course *Initiation à la recherche*
- Berg-Kirkpatrick et al. (2012) *An Empirical Investigation of Statistical Significance in NLP*
- Bruce et al. (2020) *Practical Statistics for Data Scientists*, 2nd Ed.
- Dror et al. (2018) The Hitchhiker's Guide to Testing Statistical Significance in Natural Language Processing
- Dror et al. (2020) *Statistical Significance Testing for NLP*
- Efron & Tibshirani (1994) *An introduction to the bootstrap*
- Çetinkaya-Rundel & Hardin (2022). *Modern Statistical Methods for Psychology* (esp. chapter 7)

Sources ii

- Scholivet (2021) *Analyse morpho-syntaxique massivement multilingue à l'aide de ressources typologiques, d'annotations universelles et de plongements de mots multilingues*
- Søgaard et al. (2014) *What's in a p-value in NLP?*
- Yeh (2000) *More accurate tests for the statistical significance of result differences*
- Wikipedia Statistical significance
- Youtube channels: *DATAtab*, *StatQuest*
- Discussions with Elie Antoine, Thomas Blanchard, Benoit Favre, Alexis Nasr, Shiva Taslimipoor, Marius Thorre, Abigail Walsh
- Feedback from participants of previous course editions

Sources iii

- Slides illustrated with the help of: Google images, `imgupscaler.com`, Canva
- Slides written with the help of: ChatGPT, Google Bard, DeepL, Linguee, Overleaf
- Funding: French ANR, through SELEXINI project (ANR-21-CE23-0033-01)

Backup slides

Multiple comparisons

- Multiple comparisons: probability of chance effect increases
- Bonferroni's correction

  $\rightarrow$ Divide significance level α by the number of comparisons N
- Large literature on multiple comparisons

  $\rightarrow$ Tukey's honest significance test

  $\rightarrow$ Replicability analysis (Dror et al. 2020)

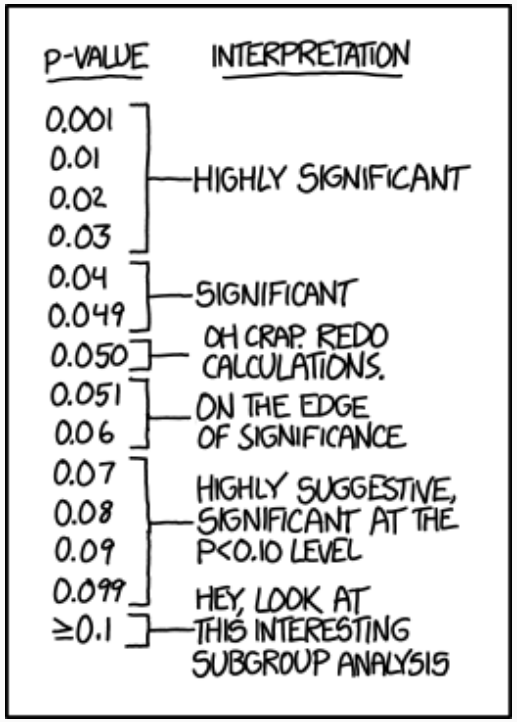
  $\rightarrow$ ...

P-hacking

A significant p-value can always be obtained

→ As long as the sample is large enough

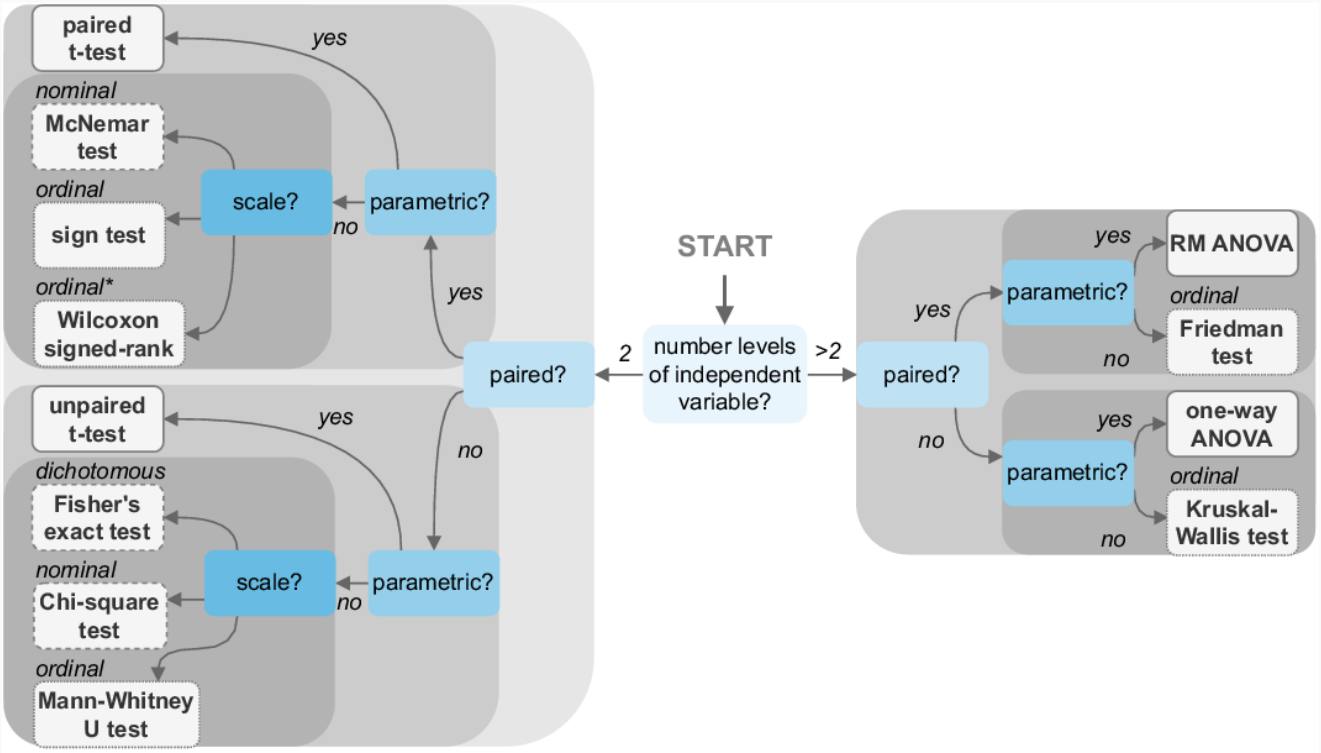
→ `https://www.youtube.com/watch?v=HDCOUXE3HMM`



Source: `https://xkcd.com/1478/`

Unpaired samples

- We only covered significance for paired samples
  - → Two systems A and B, same dataset items (x,y)
  - → Other tests for unpaired samples



Source: https://doi.org/10.1017/S1351324922000535